Comp. & Maths. with Appls. Vol. 1, pp. 79–88. Pergamon Press, 1975. Printed in Great Britain.

OPTIMAL WEIGHTING FUNCTION FOR THE INVARIANT IMBEDDING ESTIMATOR*

E. STANLEY LEE† and P. K. MISRA‡

Departments of Chemical and Electrical Engineering, University of Southern California, Los Angeles, California 90007, U.S.A.

(*Received* 11 *September* 1973)

Abstract—One of the problems in using the invariant imbedding estimator is to find the optimal or near optimal initial conditions for the weighting functions. Computational experiences indicate that these initial conditions influence the convergence rate tremendously. This problem is further complicated by the fact that the number of weighting functions increases quadratically with the number of parameters or variables to be estimated. It is not a simple matter to estimate the initial conditions to be used for a large number of interconnected weighting functions. In this work, least squares criterion combined with various optimization schemes is used to obtain the optimal initial conditions. It is shown that the convergence rate can be improved tremendously. These improved convergence rates should be very useful for off-line estimations with a limited number of experimental data.

1. THE OPTIMIZATION PROBLEM

To illustrate the problem, consider the vector differential equation

$$\frac{dX}{dt} = f(X, t) \tag{1}$$

where f and X are m dimensional vectors. Assume that only n X's or combinations of X's can be measured. Let the measured or observed quantities be

$$Z(t) = h(X, t) + \text{(observation errors)} \tag{2}$$

where $0 \leq t \leq t_f$ and h represents the measurable quantities. Both Z and h are n dimensional vectors. In general, $n \leq m$.

The problem is to estimate the values of X based on the observed values Z. The estimator equations for this problem can be obtained by the use of invariant imbedding and the least squares criterion[1, 2]. These estimator equations are

$$\frac{de}{dt} = f(e, t) + Q(t)h_c^T[Z(t) - h] \tag{3}$$

$$\frac{dQ}{dt} = f_e(e, t)Q(t) + Q(t)[f_e(e, t)]^T + Q(t)\{h_{ee}[Z(t) - h]\}Q(t) - Q(t)h_c^T h_c Q(t) \tag{4}$$

where e represents the optimal estimates for X and is an m dimensional vector, Q is the

* Supported by the National Science Foundation under Grant No. GP29049 and the Atomic Energy Commission, Division of Research under Contract No. AT(04-3)-113, Project 19.

† On leave from Kansas State University.

‡ Graduate student, Kansas State University.

weighting matrix or adjoint matrix and is an $m \times m$ dimensional matrix, and the symbols f_e and h_e represent partial differentiation of the vectors f and h with respect to e.

Equations (3) and (4) represent a system of ordinary differential equations. If the initial conditions for these equations are given or can be guessed these equations can be integrated easily on a computer to obtain the optimal estimates of X. Because of our knowledge about the physical system, approximate initial values for $e(0)$ can, in general, be guessed. However, this is not the case for the initial values of $Q(0)$. Furthermore, the initial values of this adjoint matrix, $Q(0)$, influence the convergence rate of the estimator equation tremendously. Erroneous choice of these initial conditions may cause the estimation process to diverge. This is further complicated by the fact that the number of q's increases quadratically with the number of variables. It is not a simple matter to obtain the best initial conditions for $Q(0)$.

The problem is to choose the best initial values for the weighting matrix $Q(0)$ such that the convergence rate of the estimator equations is at a maximum.

There are many ways to obtain the best initial values for the weighting matrix Q. In this work, the classical least squares criterion will be used. The problem is to find the values of the matrix $Q(t)$ at $t = 0$ such that the following expression is minimized.

$$\Omega = \|R\|^2 = \sum_{t=0}^{t^*} [Z(t) - h(e(t), t)]^T [Z(t) - h(e(t), t)] \tag{5}$$

where Ω is the residual norm. It should be emphasized that Z represents the observed values and $h(e(t), t)$ represents the optimal estimates of h. The symbol t^* represents the initial time period over which we wish to carry out the minimization. In equation (5), we have assumed that the observed values are discrete values. If continuous observations can be obtained the summation in equation (5) would be replaced by an integral.

It should be noted that equation (5) is minimized with respect to the initial values of $Q(0)$. Thus, this is an optimization problem in ordinary calculus. Any numerical search technique or even the method of implicit differentiation can be used to optimize equation (5).

In this work, two optimization techniques will be used to obtain the optimal values for the matrix $Q(0)$. The first approach is a search technique and in the second approach implicit differentiation is used to obtain the best direction to improve the objective function.

2. OPTIMIZATION BY SEARCH TECHNIQUES

To illustrate the approach, let us consider the system represented by the simple equation

$$\frac{dX}{dt} = -kX^2. \tag{6}$$

This problem has been solved in Ref. [1] and equation (6) represents a simple second order chemical reaction. The constant k is the rate constant. The invariant imbedding estimator equations for the process represented by equation (6) are

$$\frac{de(t)}{dt} = -ke^2(t) + [Z(t) - e(t)]q(t) = f_1 \tag{7}$$

$$\frac{dq(t)}{dt} = -4ke(t)q(t) - q^2(t) = f_2. \tag{8}$$

Referring to equations (1)–(4), we see that $m = n = 1$, $h(X, t) = X$, and $Q(t)$ reduces to a scalar. The problem is to find the value of $q(0)$ such that the following objective function is minimized

$$\Omega = \|R\|^2 = \sum_{t=0}^{t^*} [Z(t) - e(t)]^2. \tag{9}$$

To obtain the observed data Z computationally, equation (6) is first solved with the following numerical values

$$X(0) = 1{\cdot}0, \qquad k = 0{\cdot}05, \qquad \Delta t = 0{\cdot}1, \qquad t_f = 50$$

where Δt is the time increment between observations. Then, the value, $X(t)$, just obtained is corrupted with noise. Gaussian noise is used with zero mean and a standard deviation of unity.

In order to save computer time, the value of t^* used must be chosen with care. Obviously, the best optimal values would be obtained if $t^* = t_f$, or equation (9) is minimized over the entire time interval. However, the computer time required increases very fast with the increase of t^*. Notice that both equations (7) and (8) must be integrated to t^* in each search or movement in the search scheme. Furthermore, if the initially guessed value for $q(0)$ is not too far removed from the optimal, the inclusion of the part near t_f in the search would not significantly improve the optimal. This is due to the fact that the correct estimate of e has most probably been obtained during the initial time period. In order to find the best value of t^*, different values will be used.

Let N represent the observed data points used in the optimization scheme and Δt represents the integration step size, then

$$N = \frac{t^*}{\Delta t} + 1.$$

Any search technique can be used to minimize equation (9) with respect to $q(0)$. For simplicity in programming, the random search scheme discussed by Lee[1] is used. The results are summarized in Table 1. The $e(0)$ in Table 1 represents the initial condition assumed for e and $q(0)$ is the initial approximation assumed before any optimization search. It can be seen that the same optimal result for $q(0)$ is obtained whether $N = 100$ or $N = 50$ is used.

To compare the convergence rates in estimating e, equations (7) and (8) are solved using the optimal $q(0)$ and the solutions are plotted in Fig. 1. The results obtained by Lee[1] are also shown in Fig. 1. The symbol * is used to denote the optimal results. It can be seen that the convergence rate has improved tremendously.

Figure 2 shows the behavior of the objective function as a function of $q(0)$. The most noticeable behavior is that the objective function is fairly flat over a fairly wide range of

Table 1. Results with random search technique

N	$e(0)$	$q(0)$	Optimal $q(0)$	Optimal Ω
100	2·0	0·1	16·32	2·37
100	0·5	0·1	13·81	1·34
50	2·0	0·1	16·32	1·87
50	0·5	0·1	13·81	0·84

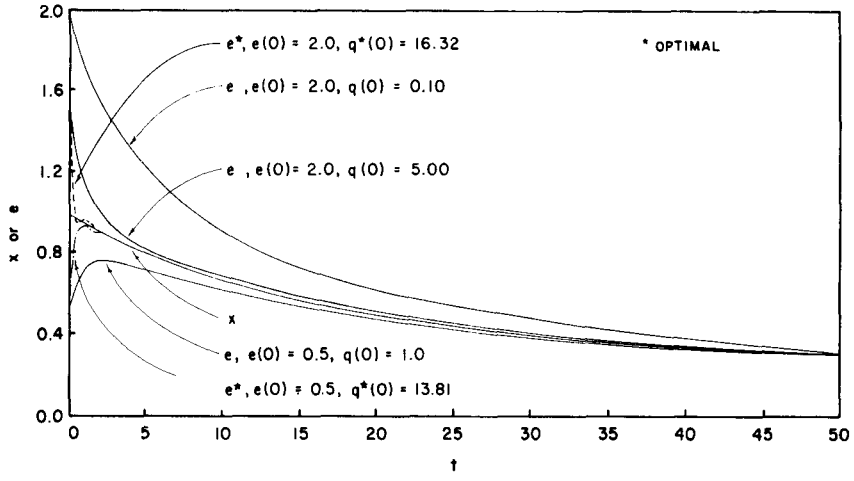

Fig. 1. Improvement in convergence rate as compared to Lee[1].

$q(0)$. Another behavior is that there exists a well-defined value of $q(0)$ over which the estimation algorithm diverges. This well-defined value is just above the optimal value of $q(0)$.

3. ESTIMATION OF STATE AND PARAMETER

Instead of only estimating X, the rate constant k can also be estimated simultaneously from the measurement on X. In addition to equation (6), we also have

$$\frac{dk}{dt} = 0. \tag{10}$$

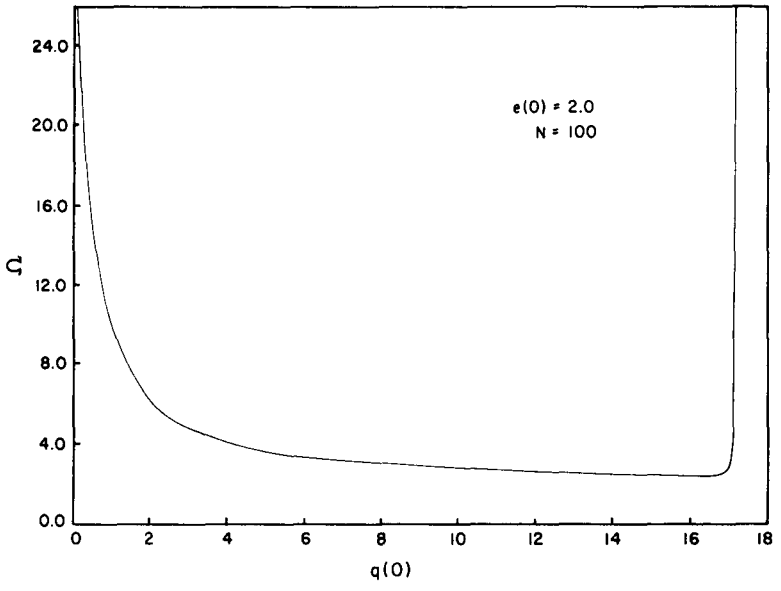

Fig. 2. Objective function as a function of $q(0)$.

Equations (6) and (10) represent the system. This estimation problem has been solved by Lee[1] with guessed initial conditions for $q(0)$. In this work, the optimal $q(0)$ will be obtained and used.

From equations (3) and (4), the invariant imbedding estimator equations are

$$\begin{bmatrix} \frac{de_1}{dt} \\ \frac{de_2}{dt} \end{bmatrix} = \begin{bmatrix} -e_2 e_1^2 \\ 0 \end{bmatrix} + \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} [Z - e_1] \tag{11}$$

$$\begin{bmatrix} \frac{dq_{11}}{dt} & \frac{dq_{12}}{dt} \\ \frac{dq_{21}}{dt} & \frac{dq_{22}}{dt} \end{bmatrix} = \begin{bmatrix} -2e_2e_1 & -e_1^2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} + \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} \times \begin{bmatrix} -2e_2e_1 & 0 \\ -e_1^2 & 0 \end{bmatrix} - \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} [1 \ 0] \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} \tag{12}$$

where e_1 and e_2 are the optimal estimates of x and k respectively; and Z is the observation on x described earlier.

Referring to equations (1)–(4), we see that $m = 2$, $n = 1$ and

$$f = \begin{bmatrix} -kX^2 \\ 0 \end{bmatrix}, \qquad h = X.$$

The problem is to find the initial values for the four dimensional matrix such that equation (9) is minimized.

The number of independent or control variables in the above problem is equal to the number of elements in the weighting matrix $Q(0)$. This number increases quadratically with the number of variables to be estimated. Thus, this number can be fairly large. Furthermore, from computational experience, it has been found that once the best initial values for the diagonal elements in the weighting matrix Q have been obtained, nearly best convergence rate for all practical purposes is also obtained provided that a reasonable set of values are used for the off-diagonal elements. In this section, only the diagonal elements, $q_{11}(0)$ and $q_{22}(0)$, will be considered for optimization and all the off-diagonal elements are assumed equal to one.

The Hooke and Jeeves[3] search technique is used to solve the optimization problem. Numerical Runge–Kutta technique is used to integrate the differential equations. The Hooke and Jeeves search technique is a sequential search technique and is fairly simple to implement. It essentially moves from one base point to another by a combination of exploratory and pattern moves. The initial $Q(0)$ guessed forms the starting base point. Failure of a pattern move causes the search to return to its old base point. Step size used in the search is reduced when all exploratory moves fail. The procedure is repeated until the step size is reduced to a prespecified accuracy.

The computational results are summarized in Table 2. The observed data points and the various numerical values used are the same as those discussed earlier. A step size of 5 was used in Hooke and Jeeves search technique. The minimum step size allowed in this search was 1·0.

To compare the convergence rates in estimating k, the results obtained by Lee[1] are

Table 2. Results with Hooke–Jeeves search technique

					Optimal		
N	$e_1(0)$	$e_2(0)$	$q_{11}(0)$	$q_{22}(0)$	$q_{11}(0)$	$q_{22}(0)$	Optimal Ω
100	0·5	0·1	6	6	14·75	6·00	1·393
100	0·5	0·1	5	5	15·00	5·00	1·391
100	2	0·1	5	5	15·00	22·50	2·408
120	2	0·1	6	6	14·12	32·25	2·686

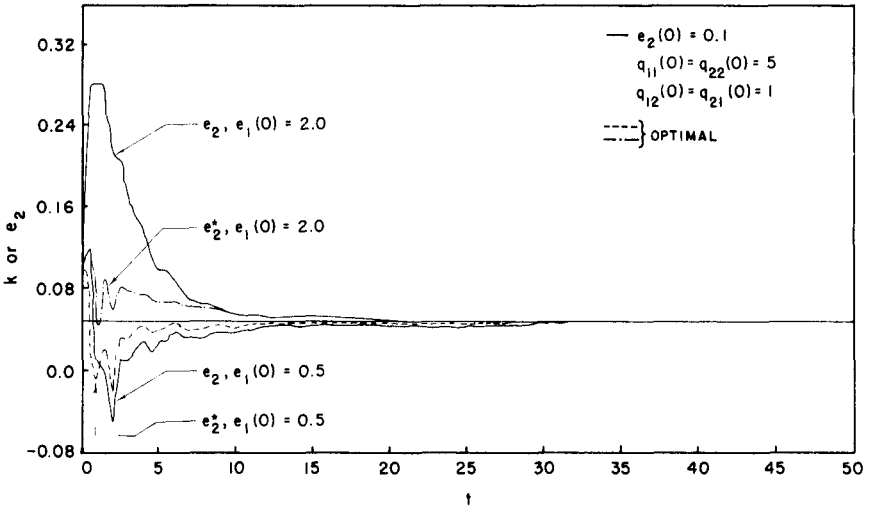

Fig. 3. Improvement in convergence rate in estimating k as compared to Lee[1].

shown in Fig. 3. The convergence rates for estimating k by using the optimal values of $q_{11}(0)$ and $q_{22}(0)$ are also shown in Fig. 3. For $e_1(0) = 2$, the optimal values are

$$q_{11}(0) = 14{\cdot}12, \qquad q_{22}(0) = 32{\cdot}25$$

and for $e_1(0) = 0{\cdot}5$, the optimal values are

$$q_{11}(0) = 15 \quad \text{and} \quad q_{22}(0) = 5.$$

The convergence rates for estimating X are not shown. However, the results are fairly similar to those shown in Fig. 1.

It should be noted that the values of N or t^* used influences the optimal value obtained for $e_1(0) = 2$. It was found that with $N = 120$ or $t^* = 12$, a much better optimal was obtained.

4. OPTIMIZATION BY A GRADIANT TECHNIQUE

Since the minimization of equation (5) is an optimization problem in ordinary calculus, many other optimization techniques can be used. In this section, a gradient technique will be developed and applied to the minimization of equation (5). The technique essentially consists of a scheme for the iterative improvement of the objective function in the gradient

direction. However, since equations (3) and (4) are differential equations, implicit differentiation is used to obtain the gradient direction.

In order to save computer time, equation (5) will be minimized with respect to only the diagonal elements in the adjoint matrix $Q(0)$. This choice seems partially justifiable from actual computational experiences which indicate that in general a proper choice of the diagonal elements is enough to obtain a fast convergence rate. For all practical purposes, this obtained fast convergence rate is nearly the best convergence rate provided that a reasonable set of values is used for the off-diagonal elements. Thus, define the control vector π as

$$\pi = [q_{11}(0), q_{22}(0), \ldots, q_{mm}]. \tag{13}$$

In order to minimize equation (5) with respect to π, the direction where the most rapid decrease in Ω occurs will be used. This is the gradient direction which can be obtained by differentiating equation (5) with respect to π

$$\frac{d\Omega}{d\pi} = 2\left[\frac{dR}{d\pi}\right]^T R(\pi). \tag{14}$$

Now, let us define the relationship

$$\pi^{k+1} = \pi^k - \gamma_k^* \left[\left(\frac{dR}{d\pi}\right)^k\right]^T R(\pi^k) \tag{15}$$

where

$$\gamma_k^* = \gamma_k \frac{\|R^k\|^2}{\left\|\left[\left(\frac{dR}{d\pi}\right)^k\right]^T R^k\right\|^2} \tag{16}$$

where k represents the iteration number and γ_k represents the step size or the distance to move in the gradient direction.

With the values of π at the kth iteration known, improved values for π at the $(k+1)$st iteration can be obtained by using equation (15). This procedure can be continued until the norm $\|R\|^2$ has reduced to a prespecified value.

Since equations (3) and (4) are differential equations, the expression for $dR/d\pi$ cannot be obtained easily. To simplify the manipulations, let us consider the problem represented by equations (6)–(9). For this problem equation (13) reduces to

$$\pi = q(0) \tag{17}$$

and equation (14) becomes

$$\frac{d\Omega}{d\pi} = 2\frac{dR}{d\pi} R(\pi). \tag{18}$$

Using implicit differentiation, we have

$$\frac{dR}{d\pi} = \frac{1}{2}\left[\sum_{t=0}^{t^*} \{e(t) - Z(t)\}^2\right]^{1/2}\left[\sum_{t=0}^{t^*} \left\{\frac{\partial h}{\partial e}\bigg|_t \frac{\partial e(t)}{\partial \pi} + \frac{\partial h}{\partial q}\bigg|_t \frac{\partial q(t)}{\partial \pi}\right\}\right] \tag{19}$$

where

$$h = [e(t) - Z(t)]^2. \tag{20}$$

Some of the terms in the above can be reduced to

$$\left.\frac{\partial h}{\partial e}\right|_t = 2[e(t) - Z(t)] \tag{21}$$

$$\left.\frac{\partial h}{\partial q}\right|_t = 0. \tag{22}$$

Now, expressions for $\partial e/\partial \pi$ and $\partial q/\partial \pi$ must be obtained. In order to accomplish this, integrate equations (7) and (8)

$$e(t) = e(0) + \int_0^t f_1 \, dt \tag{23}$$

$$q(t) = q(0) + \int_0^t f_2 \, dt. \tag{24}$$

Partial differentiating both sides of equation (23) with respect to π, we obtain

$$\frac{\partial e(t)}{\partial \pi} = 0 + \int_0^t \left[\left.\frac{\partial f_1}{\partial e}\right|_t \frac{\partial e(t)}{\partial \pi} + \left.\frac{\partial f_1}{\partial q}\right|_t \frac{\partial q(t)}{\partial \pi}\right] dt. \tag{25}$$

Differentiating the above equation with respect to time, we obtain

$$\frac{d}{dt}\left[\frac{\partial e(t)}{\partial \pi}\right] = \frac{\partial f_1}{\partial e(t)}\left[\frac{\partial e(t)}{\partial \pi}\right] + \frac{\partial f_1}{\partial q(t)}\left[\frac{\partial q(t)}{\partial \pi}\right]. \tag{26}$$

Similarly, from equation (24) we obtain

$$\frac{d}{dt}\left[\frac{\partial q(t)}{\partial \pi}\right] = \frac{\partial f_2}{\partial e(t)}\left[\frac{\partial e(t)}{\partial \pi}\right] + \frac{\partial f_2}{\partial q(t)}\left[\frac{\partial q(t)}{\partial \pi}\right]. \tag{27}$$

To simplify notation, let us define

$$X_1 = \frac{\partial e(t)}{\partial \pi}, \qquad X_2 = \frac{\partial q(t)}{\partial \pi}. \tag{28}$$

Equations (26) and (27) can be rewritten as

$$\frac{dX_1}{dt} = X_1 \frac{\partial f_1}{\partial e(t)} + X_2 \frac{\partial f_1}{\partial q(t)} \tag{29}$$

$$\frac{dX_2}{dt} = X_1 \frac{\partial f_2}{\partial e(t)} + X_2 \frac{\partial f_2}{\partial q(t)}. \tag{30}$$

Initial conditions for equations (29) and (30) can be obtained from the definitions in equation (28). At the initial time $t = 0$, we have

$$X_1(0) = \frac{\partial e(0)}{\partial \pi} = 0, \qquad X_2(0) = \frac{\partial q(0)}{\partial \pi} = 1. \tag{31}$$

Performing the partial differentiations indicated, equations (29) and (30) become

$$\frac{dX_1}{dt} = (-2ke - q)X_1 + (Z - e)X_2 \tag{32}$$

$$\frac{dX_2}{dt} = (-4kq)X_1 + (-4ke - 2q)X_3. \tag{33}$$

Using the initial conditions given in equation (31), $X_1(t)$ and $X_2(t)$ can be obtained easily by integrating equations (32) and (33). The results from this integration can be substituted into equation (19) and thus the differential $dR/d\pi$ can be obtained.

The computational procedure can now be summarized as follows:

(1) Guess an initial approximation for the control $q(0)$ or π, call this initial guess π^0;
(2) Integrate equations (7) and (8) using π^0 as the initial condition for q. The initial condition for $e(0)$ can be obtained from our knowledge concerning the process;
(3) Integrate equations (32) and (33) using the initial condition given in equation (31);
(4) Compute the norm using equation (9) and $dR/d\pi$ using equations (19)–(22);
(5) Choose a suitable step size, γ_k, and compute an improved π by using equations (15) and (16);
(6) Repeat steps (2)–(5) until the norm $\|R\|^2$ has been reduced to a prespecified accuracy.

The problem represented by equations (6)–(9) was solved by the gradient technique. The same numerical values used previously in the search technique are used here. The other numerical values used are

$$e(0) = 2, \qquad \pi^0 = q(0) = 1, \qquad \gamma_0 = 0{\cdot}5, \qquad t^* = 10.$$

Note that only the first part of the time interval, namely, $(t_0, t^*) = (0, 10)$, is used in the optimization calculations. The optimization calculations are terminated and the optimal values are considered obtained when

$$|\Omega^{k+1} - \Omega^k| = 0{\cdot}001. \tag{34}$$

The convergence rate of the gradient technique is shown in Table 3. It can be seen from Table 3 that a fairly good estimate for $q(0)$ was obtained in three to five iterations. Figure 4 illustrates the convergence rates for estimating X with $q(0)$ values obtained from

Table 3. Convergence rate of $q(0)$ by the gradient approach

Iteration (k)	$\pi^k[= q(o)]$	γ^k	Ω^k	$\frac{dR}{d\pi}$
0	1·0	0·5	9·948	−1·077
1	2·464	0·5	5·433	−0·306
2	6·273	0·5	3·311	−0·0647
3	13·307	0·25	2·522	−0·0232
4	15·442	0·031	2·374	−0·0069
5	16·318	0·0039	2·373	0·0064
6	1·349	0·0625	8·111	−0·732
7	3·295	0·5	4·557	−0·193
8	8·825	0·5	2·918	−0·0379
9	14·465	0·125	2·433	−0·0164
10	15·953	0·0156	2·360	0·00006

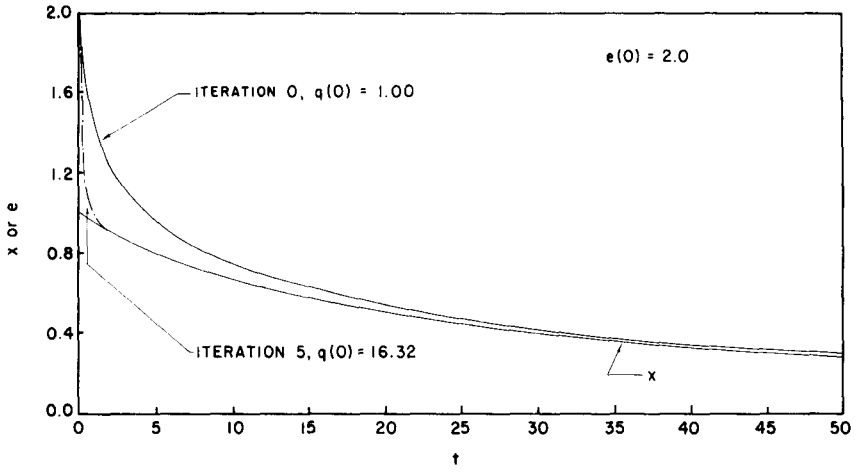

Fig. 4. Convergence rates for estimating x with different iteration results from the gradient optimization.

different gradient iterations. It is interesting to note that the optimal value of $q(0)$ obtained in the fifth iteration coincides with that obtained by the search procedure previously.

REFERENCES

1. E. S. Lee, *Quasilinearization and Invariant Imbedding*, Academic Press, New York (1968).
2. R. Bellman, H. H. Kagiwada, R. Kalaba and R. Sridhar, Invariant Imbedding and Nonlinear Filtering Theory, RM-4374-PR, RAND Corporation, Santa Monica, California (1964).
3. R. Hooke and T. A. Jeeves, Direct search solution of numerical and statistical problems, *J. Assoc. Computing Machinery* **8,** 212 (1961).